



Современные программные решения для преобразования геномных координат между различными версиями сборок генома человека

Бобрик П.Ю., Вертёлко В.Р., Гурьянова И.Е.

Республиканский научно-практический центр детской онкологии, гематологии и иммунологии, Республика Беларусь

Аннотация

В данной работе исследуются современные программные решения для преобразования геномных координат между различными версиями сборок генома человека. Необходимость использования актуальных версий референсных геномов во многом связана с ограничениями, вызванными использованием устаревших сборок.

Исследование представляет обзор существующих программ для преобразования геномных координат и призвано помочь исследователям выбрать подходящий инструмент в зависимости от их потребностей и доступных ресурсов.

Введение

Референсная геномная сборка является необходимым ресурсом при работе с данными высокопроизводительного секвенирования. На сегодняшний день существует две актуальные версии геномных сборок человека — GRCh37 и GRCh38. Самой свежей считается версия GRCh38, одним из основных обновлений в которой стало закрытие многочисленных пробелов, присутствующих в предыдущей сборке. Несмотря на это, многие исследователи не решаются перейти на актуальную версию геномной сборки, так как подобный переход может потребовать значительных изменений в алгоритме обработки данных. В свою очередь, использование GRCh37 накладывает ряд ограничений на возможность работы с актуальными базами данных. Нами была поставлена цель провести обзор существующих программных решений, позволяющих осуществлять преобразование геномных координат между различными версиями сборок генома человека в файлах формата vcf.

Материалы и методы

Были протестированы следующие биоинформатические инструменты для преобразования геномных координат: *CrossMap*, *LiftoverVcf* (Picard), *transanno*, *Assembly Converter* (Ensembl). Для оценки работоспособности программ использовались 10 VCF-файлов ($\approx 1,5$ Гб каждый), содержащих в среднем 4,5 млн. замен, выявленных в ходе полногеномного секвенирования, а также файл GRCh37_to_GRCh38.chain, необходимый для выполнения конвертации. Тестирование осуществлялось в ОС Ubuntu (22.04).

Результаты и обсуждение

Программа CrossMap смогла обработать до 98,70% всех замен в среднем за 2 минуты 17 секунд (± 10 сек.) с минимальным использованием вычислительных ресурсов. Инструмент transanno показал схожие результаты: было преобразовано 98,72% вариантов за 2 минуты 7 секунд (± 6 сек.). Программа LiftoverVcf (Picard) конвертировала 98,67% замен в среднем за 1 минуту 15 секунд (± 5 сек.), однако потребление оперативной памяти в пиковый момент достигало 12,8 Гб. Веб-инструмент Assembly Converter (Ensembl) даёт возможность запустить преобразование координат в удобном графическом интерфейсе. Стоит отметить, что максимальный размер загружаемого в Assembly Converter файла не должен превышать 50 Мб, что накладывает ограничение на работу с данными полноэкзомного и полногеномного секвенирования.

Таблица 1. Функциональные характеристики тестируемых программ

	efficiency, %	t, sec.	RAM, Gb
CrossMap	98,70	137 $\pm$ 10	0.4
transanno	98,72	127 $\pm$ 6	0.4
LiftoverVcf	98,67	75 $\pm$ 5	12.8
Assembly Converter	50 Mb limit		

Заключение

Было рассмотрено 4 инструмента для преобразования геномных координат. CrossMap и transanno можно назвать наиболее универсальными вариантами для данной задачи, т.к. данные программы легко устанавливаются и не требуют больших вычислительных мощностей. Assembly Converter является хорошим вариантом для специалистов с небольшим опытом работы в командной строке. LiftoverVcf (Picard) удобен в случаях, когда нужно провести преобразование в более сжатые сроки за счёт дополнительных вычислительных ресурсов.

Контакты:

Бобрик Павел Юрьевич, м.н.с.
Email: bobrykpavel2001@gmail.com

Вертелко Владислав Русланович, м.н.с.
Email: vertelko.vladislav@gmail.com

